Policy, Planning, and Research

WORKING PAPERS

Education and Employment

Population and Human Resources Department
The World Bank
July 1989
WPS 242

A Multi-Level Model of School Effectiveness in a Developing Country

Marlaine E. Lockheed
and
Nicholas T. Longford

Schools in Thailand are more uniformly effective than previous research in developing countries would suggest. Higher levels of math achievement are associated with more qualified math teachers, an enriched curriculum, and frequent use of textbooks.

Policy, Planning, and Research

WORKING PAPERS

Education and Employment

What makes one school more effective than another — particularly which inputs and management practices most efficiently enhance student achievement — has become the center of lively debate in the literature. Which method to use to compare school effects particularly concerns analysts.

Lockheed and Longford used a multi-level model to analyze what improved performance in grade 8 mathematics in Thailand. They concluded that:

- Schools in Thailand were equally effective in teaching students eighth grade mathematics (for example, in transforming pretest scores into posttest scores).

- Schools and classrooms contributed 32 percent of the variance in posttest scores and individual characteristics 68 percent.

- Greater learning occurred in schools having a higher proportion of teachers qualified to teach mathematics, classrooms having an enriched curriculum and in which textbooks frequently were used.

- Learning was higher for boys, younger students, and for children who reported higher educational aspirations, less parental encouragement, more confidence in their own mathematics ability, greater interest in mathematics, and a feeling that mathematics was relevant to them.

- Schools in Thailand were more uniform in their effects on learning than previous research in developed countries had suggested would be the case.

The model developed by Lockheed and Longford was able to explain most variance between schools but significantly less within schools. Only one variable slope was observed: the relationship between educational aspirations and achievement.

Lockheed and Longford applied multi-level techniques to longitudinal data recently collected by the International Association for the Evaluation of Educational Achievement in Thailand.

One question they tried to answer was: How do estimates obtained from the new multi-level techniques compare with those obtained from ordinary regression methods?

This paper is a product of the Education and Employment Division, Population and Human Resources Department. Copies are available free from the World Bank, 1818 H Street NW, Washington DC 20433. Please contact Cynthia Cristobal, room S6-001, extension 33640 (66 pages with tables).

Acknowledgements

The contributions of the International Association for the Evaluation of Educational Achievement (IEA) in making the data available to us and the extensive comments of Stephen Raudenbush on an earlier draft are gratefully acknowledged.

CONTENTS

Tables

A MULTI-LEVEL MODEL OF SCHOOL EFFECTIVENESS IN A DEVELOPING COUNTRY

INTRODUCTION

There are several central questions behind the research into school effectiveness. First, do schools make a difference in how much a student learns (that is, does the specific school in which a child is enrolled have a particular impact on his or her achievement, independent of family background)? Second, if so, what are the characteristics of the school that account for this difference? Third, do certain schools affect certain types of students differently than others?

These questions, first raised by Coleman in the 1960s, have been reconsidered in the current research on the effectiveness of private schools (Coleman, Hoffer and Kilgore 1982) and by a new generation of "effective school" researchers (Aitkin and Longford 1986; Goldstein 1986; Raudenbush and Bryk 1986; Reynolds 1985; Rutter 1983; Willms 1987). The new researchers have investigated the questions through the application of new analytic techniques that take into account the hierarchical nature of most data on education: children within classrooms, classrooms within schools and schools within educational authorities (e.g., districts).

Although appropriate methods for analyzing hierarchically structured data on education have been available since the early 1970s (Dempster, Laird and Rubin 1977; Lindley and Smith 1972), application of these methods to educational policy decisions in developing countries has been hampered by two important shortcomings: (i) the absence of computationally efficient algorithms for multi-level analysis; and (ii) the lack of adequate data (sufficient cases at each organizational level). Recently, new

computational methods have been developed that address the first problem (Goldstein 1984, 1986; Longford 1987; Bryk, Raudenbush, Seltzer and Congdon, Jr. 1986), and data sets sufficient for their application have been collected in a number of developing countries.

This paper applies multi-level techniques to longitudinal data recently collected by the International Association for the Assessment of Educational Achievement (IEA) in Thailand to answer the following questions: (i) do Thai middle schools affect student learning differentially? (ii) what part of the variation in student learning is attributable to between school characteristics versus between student characteristics? (iii) what characteristics of teachers and schools enhance student achievement, independent of student background? (iv) what is the comparative effectiveness of alternative school inputs? (v) are the effects of schools uniform across different students? and (vi) how do estimates obtained from the new, multi-level techniques compare with those obtained from ordinary regression methods?

Background

The comparative effectiveness of schools in developing countries, particularly the relative efficiency with which alternative inputs and management practices enhance student achievement, has become the center of a lively debate in the literature (see, for example, Fuller 1987; Harbison and Hanushek 1989; Heyneman 1986; Lockheed and Hanushek 1988). These issues have important implications for how governments and international development agencies should allocate their limited resources--whether they should concentrate on certain types of inputs (capital investment or lowering class size) or should finance others (instructional materials, teacher or headmaster

training or student testing). In the United States and the United Kingdom, the debate was sparked by studies that claimed to identify effective schools: those that enhanced student achievement more than other schools working with similar students and material inputs (see Raudenbush 1987 for a recent review).

In developing countries, research on school effectiveness has been more limited, and studies examining the effects of alternative inputs on student achievement have not taken into account the explicitly hierarchical nature of the explanatory models and data. Instead, most research on effective schools in developing countries has utilized a "production function" approach that compares the relative effectiveness of alternative material and non-material inputs and, to a lesser degree, teaching processes on student achievement. The school characteristics most frequently examined have been indicators of material inputs: per pupil expenditures, number of books, presence of a library, presence of desks, teacher salaries and so forth.[1/] The past decade has provided several important reviews of this research (Avalos and Haddad 1981; Fuller 1987; Heyneman and Loxley 1983; Husen, Saha and Noonan 1978; Schiefelbein and Simmons 1981; Simmons and Alexander, 1978). Most of the reviews conclude that, when student background is controlled for, school characteristics do have significant effects on achievement, and, in many cases, the effects of school characteristics are greater than the effects of family background.

[1/] The most extensive research using this type of model is reported in a recent longitudinal study (Harbison and Hanushek 1989) of the effects of material inputs on student achievement in rural Brazil.

Heyneman and Loxley (1983), for example, found that the variance in student achievement explained by three family background variables averaged 8.6% across 17 developing countries, while the variance explained by school characteristics amounted to 16%, nearly twice as great. Yet, overall, the amount of variance in student achievement explained by variables related to family background and school inputs in developing countries remains remarkably low in comparison with the results of similar studies conducted in developed countries. Heyneman (1986) has argued strongly that the failure of conventional models to explain the variance in achievement is a consequence of poorly conducted research. An equally strong case can be made regarding the inadequacy of the models and indicators employed.

The more recent research on school effectiveness differs from earlier approaches in four important ways. First, education production function research has moved away from answering the questions of whether and how much specific material and non-material inputs affect student achievement to exploring other questions, including the effects of alternative inputs on achievement (e.g., Harbison and Hanushek 1989) and the mechanisms whereby material and non-material inputs affect achievement (Lockheed, Vail and Fuller 1987). Second, better and more culturally relevant indicators of students' social background in developing countries have been utilized (e.g., Lockheed, Fuller and Nyirongo 1987). Third, complex organizational models of student achievement (e.g., Rosenholtz 1989) have begun to replace education production function models. Fourth, research has begun to center on the classroom and classroom processes as important determinants of learning, with specific focus on the role of teachers and administrators as managers of student learning

(e.g., Lockheed and Komenan 1989; Lockheed, Fonacier and Bianchi 1989). This paper addresses all four issues.

Methodological Considerations

While matters of substantive concern continue to drive the research on effective schools, the "effective schools" issue has been fueled by controversy over statistical methodology, interpretation and data (for example, Sirotnik and Burstein 1985). The most important statistical issue is the use of appropriate methods to analyze multi-level data. The argument concerns how behavior at one level (e.g., classroom, school or district) influences behavior at a different level (e.g., students) and how to estimate these multi-level effects correctly.[2/]

Hierarchically structured data are common in social research, because social institutions are typically hierarchically organized. However, the commonly used statistical techniques for dealing with related data may lead to biased estimates.[3/] In particular, it has been established that, when observations within clusters on any stratum are more homogeneous than those between clusters, the use of ordinary regression methods (e.g., OLS) with such data can lead to biased estimates of regression coefficients in unbalanced designs and even to substantially biased standard errors for these estimates in balanced designs. In that most policy research entails the use of

---

[2/] These hierarchical structures result from design elements (stratified sampling), data collection technicalities (e.g., interviewer effect) or intrinsic interest in cross-level effects (e.g., the effects of post-natal feeding programs on the relationship between birth weight and subsequent cognitive development).

[3/] An extended discussion of this issue is provided by Goldstein (1987).

unbalanced designs, a serious problem may arise when ordinary least squares regression estimates are used to quantify effects.

Proper analysis of multi-level data requires two distinct changes in thinking about the data. First, the researcher must confront the demands of the inherently hierarchical data common to education at the stage of sample design, so that sufficient numbers of units at each level are sampled (e.g., adequate samples of schools and classrooms, in addition to the sample of students). Second, and more important, hierarchical analysis allows a major shift in how the effects of organizations on individuals may be viewed: instead of considering only the effects of organizational characteristics on organizational means, the effects on relationships are also modelled. For example, certain school or classroom interventions may affect not only average student achievement, but they may also lessen the degree of association between family background and student achievement. Here an organization-level force serves to mediate an individual-level effect.

Until recently, most discussions of multi-level analysis have remained theoretical, bounded by the costs and computational requirements of existing analytic tools. However, the recent development of new analytic tools for analyzing multi-level data has energized the debate (Aitkin and Longford 1986; Goldstein 1986; Mason, Wong and Entwisle 1984; and Raudenbush and Bryk 1986). The development of the general EM algorithm (Dempster, Laird and Rubin 1977) provided a theoretically satisfactory and computationally manageable approach to estimation of covariance components in hierarchical linear models.

To date, application of these methods in education policy research has been limited to a relatively few studies of schools in developed countries. To the best of the authors' knowledge, the present study is the first such application to data from developing countries.

CHAPTER I: THE DATA

Context

The data used in this study come from the IEA Second International Mathematics Study (SIMS) in Thailand, 1981-82, and address eighth grade mathematics achievement. The structure of Thailand's education system includes six primary school grades, three lower secondary school grades, three upper secondary school grades and tertiary education. While the first six years of schooling are compulsory, secondary education is not. At the time the data were collected, 33% of the 14-year-old age cohort were enrolled in grade eight.

Sample

The IEA SIMS sample consisted of 99 mathematics teachers and their 4,030 eighth-grade students. It was derived from a two-stage, stratified random sample of classrooms. The 13 primary sampling units were the 12 national educational regions of Thailand plus the capital, Bangkok. Within each region, a random sample of lower secondary schools was selected. At the second stage, a random sample of one class per school was selected from a list of all eighth-grade mathematics classes within the school; only students

enrolled in school for the entire school year were included. The result was a 1% sample of eighth-grade mathematics classrooms within each region. This design does not distinguish between the school and classroom levels, so that only inferences about the aggregate of these effects are possible.

Method

At both the beginning and end of the school year, students were administered a mathematics test covering five content areas of the curriculum (arithmetic, algebra, geometry, statistics and measurement). Students also completed a short background questionnaire at the pretest and a longer one at the posttest administration. Teachers completed several instruments at the posttest, including a questionnaire on their background and one on general classroom processes. They also provided information about teaching practices and characteristics of their randomly selected "target" class. A school administrator provided data about the school.

Measures

The measures included indicators of student attitude and achievement, of student social class background, of material and non-material inputs at the school and classroom levels, and of classroom organization and teaching practices. The following sections provide a description of each of the variables analyzed in this paper (see Lockheed, Vail and Fuller 1987 for an extended discussion); acronyms for the variables are given in parentheses. For easier orientation, the acronyms for pupil-level variables are given in capital letters and for group-level (region/school/classroom) variables in underlined lower-case letters. This distinction will be clear from Tables 1

and 2, which provide the definitions and summary statistics for all the variables in the original data set and the data set developed as part of this paper.

Mathematics achievement. The IEA developed five mathematics tests for use in/SIMS. One of the tests was a 40-item instrument called the core test. The remaining 4 tests were 35-item instruments called rotated forms, designated A through D. The 5 test instruments contained roughly equal proportions of items from each of the 5 areas of curriculum content, except that the core test contained no statistics items. For purposes of this analysis, we regard the instruments as parallel forms with respect to mathematics content.

The IEA longitudinal design called for students to be administered both the core test and one rotated form chosen at random at both the pretest and posttest. In Thailand, students were pretested using the core test and one rotated form. At the posttest, they again took the core test and one rotated form that was different from the rotated form taken at the pretest. Approximately equal numbers of students took each of the rotated forms test in both test administrations.

One goal of this analysis was to predict posttest achievement as a function of pretest performance and other determinants. Since students took the core test during the pretest, their posttest scores would reflect, to some degree, familiarity with the test items. For purposes of our study, instead of using the core test, we analyze the scores obtained from the rotated forms, after equating them to adjust for the differences in test length and difficulty. In this analysis, we use equated rotated form formula scores for

both the pretest (XROT) and posttest (YROT) measures of student achievement in mathematics.[4/]

Table 1: Sample Characteristics and Variable Names, Descriptions and Means (Proportions) of Student-Level Variables for Three Data Sets

Variable Name	Description	Means/Proportions Data Set 1	Data Set 2	Data Set 3
Sample				
	Students	2,076	2,804	3,025
	Classrooms	60	80	86
Student-Level Variables				
XROT	Pretest mathematics achievement score	9.15	8.83	8.83
XSEX	Student gender (0 = female; 1 = male)	.53	.53	.53
XAGE	Age in months	170.94	171.05	171.09
YFOCCI	Father's occupational status:			
	Unskilled or semi-skilled worker	.15	.15	.15
	Skilled worker	.44	.45	.46
	Clerical or sales worker	.26	.26	.25
	Professional or managerial worker	.15	.15	.14
YMEDUC	Mother's educational attainment			
	Very little or no schooling	.26	.26	.26
	Primary school	.58	.58	.58
	Secondary school	.09	.09	.09
	College, university or some form of tertiary ed.	.07	.07	.06
YHLANG	Use of language of instruction at home (0 = no, 1 = yes)	.49	-	-
YHCALC	Calculator at home (0 = no, 1 = yes)	.31	-	-
YMOREED	Educational expectations			
	Less than two years	.08	.08	.08
	Two to four years	.30	.31	.30
	Five to seven years	.41	.41	.41
	Eight or more years	.22	.20	.21
YPARENC	Parental encouragement (1 = high)	2.12	2.10	2.09
YPERCEV	Perceived mathematics ability (1 = high)	4.05	4.05	4.05
YFUTURE	Perceived future importance of mathematics (1 = low)	2.06	2.05	2.06
YDESIRE	Motivation to succeed in mathematics (1 = low)	5.47	5.47	5.47

[4/] For more detail on the construction of the achievement measures, see Lockheed, Vail and Fuller (1986).

Table 2: Sample Characteristics and Names, Descriptions and Means (Proportions) of Group-Level Variables for Three Data Sets

Variable Name	Description	Means/Proportions Data Set 1	Data Set 2	Data Set 3
Sample				
Students		2,076	2,804	3,025
Classrooms		60	80	86
Group-level Variables				
senrolt	Number of students in school ('000)	1.27	1.44	1.41
sdaysyr	Days in school year	195.04	-	-
sputear	Pupil/teacher ratio in school	14.86	15.81	15.93
squalmt	% of teachers in school qualified to teach math.	.57	.62	.62
spci81	District per capita income (in 1000 bahts)	12.94	12.97	-
sstream	Ability groupings for instruction (0 = no; 1 = yes)	.46	.47	-
tsex	Teacher gender (0 = female, 1 = male)	.33	.37	-
tage	Teacher age in years	29.04	-	-
texptch	Years of teaching experience	7.25	-	-
tedmath	Semesters of post-secondary mathematics	3.95	-	-
tnstuds	Number of students in target class	43.61	42.61	-
tmthsub	Math curriculum (0 = remedial or normal, 1 = enriched)	.22	.20	.18
txtbk	Frequency of use of textbook (0 = no; 1 = yes)	.55	.56	.58
cefeed	Frequency of individual feedback	2.15	-	-
tadminl	Minutes spent weekly on routine administration	26.84	-	-
torderl	Minutes spent weekly maintaining class order	19.40	20.27	20.33
tseatl	Minutes students spent weekly at seat or blackboard	53.76	54.57	-
tvismat	Use of commercial visual materials (0 = no; 1 = yes)	.34	.40	-
tworkbk	Use of published workbooks (0 = no; 1 = yes)	.85	.83	.81

Student background characteristics. The basic background information about each student included his or her gender (XSEX), age in months (XAGE), paternal occupational status (YFOCCI), highest maternal education (YMEDUC), home language (YHLANG) and home use of a four-function calculator (YHCALC). Paternal occupation (YFOCCI) was classified into four categories: (i) unskilled or semi-skilled worker, (ii) skilled worker, (iii) clerical or sales worker, and (iv) professional or managerial worker. Maternal education (YMEDUC) was classified into four categories: (i) very little or no schooling, (ii) primary school, (iii) secondary school, and (iv) college, university or some form of tertiary education.

Student attitudes and perceptions. Five indices of student attitudes and perceptions were included. Student educational expectations (YMOREED) were measured by a single item that asked about the number of years of full-time education the student expected to complete after the current academic year. The following categories were defined: (i) less than two years, (ii) two to four years, (iii) five to seven years, and (iv) eight or more years. Parental encouragement (YPARENC) was measured by a four-item index composed of responses on a Likert-type scale in which students described their parents' interest in, and encouragement for, mathematics achievement. For example, for the item "My parents encourage me to learn as much mathematics as possible," the response alternatives ranged from "exactly like" the student's parents (= 1) to "Not at all like" the student's parents (= 5). The four items comprised a single factor, with principal component factor loadings ranging from .72 to .83 and communality of 2.43. A low score represented greater parental support. Perceived mathematics ability (YPERCEV), perceived usefulness of mathematics

(YFUTURE) and motivation toward mathematics achievement (YDESIRE) were all developed from a factor analysis of the student attitude survey, which contained Likert-type items having response alternatives ranging from "strongly disagree" (= 1) to "strongly agree" (= 5). The factors were initially identified through varimax factor analyses and then confirmed through principal component analyses, from which the factor scores were constructed. For YPERCEV, a low value represented a positive attitude; for YFUTURE and YDESIRE, a high value represented a positive attitude.

School characteristics. This study looks at data on six school characteristics. Five are conventional indicators of material and non-material inputs: (i) school size in terms of the total number of students enrolled (senrolt), an indicator of potential resources; (ii) length of the school year in days (sdaysyr), an indicator of the time available for instruction; (iii) student/teacher ratio in the school (sputear), an indicator of the availability of teacher resources for the student; (iv) percentage of the teaching staff qualified to teach mathematics (squalmt), an indicator of the quality of teacher resources; and (v) per capita income in 1981 at the district level (spci81), another indicator of resources. One measure of school organization is included: (vi) presence of ability grouping (sstream).

Teacher characteristics. Four teacher characteristics are analyzed: (i) gender (tsex); (ii) age (tage); (iii) teaching experience (texptch); and (iv) number of semesters of post-secondary mathematics education (tedmath). The latter two variables are conventional indicators of teacher quality.

Classroom characteristics. Three characteristics of the classroom are analyzed: (i) class size (tnstuds), an indicator of the teacher resources available to the student in his/her mathematics class; (ii) remedial or

typical versus enriched mathematics subject matter (tmthsub), an indicator of the quality of the curriculum for the student in a particular class; and (iii) whether or not the teacher used textbooks frequently in the class (txtbk), an indicator of the availability of instructional materials in the classroom.

Teaching practices. Six variables referring to teaching practices are considered: (i) providing feedback to students (cefeed), a composite index of five elements of teaching practice: commenting on student work, reviewing tests, correcting false statements, praising correct statements and giving individual feedback; (ii) number of minutes per week the teacher spent on routine administration (tadminl); (iii) maintaining class order (torderl); (iv) monitoring assigned seatwork (tseatl); (v) using commercially produced visual materials (tvismat); and (vi) using workbooks (tworkbk). All information on variables related to teaching practices were self-reported.

In summary, the data set contains information on 32 variables about 4,030 pupils from 99 schools. Of the 32 variables, 13 involve student characteristics, 5 refer to the school, 4 to the teacher, 9 relate to the classroom, and 1 is a characteristic of the district (catchment area). The distinction between the variables related to pupils and to classrooms/teachers/schools (henceforth called groups, since they are confounded in the design) is important because they play different roles in explaining variations in achievement.[5/]

[5/] It should be noted that the complete data set consists of 13*4,030 + 19*99 = 54,271 units of data, although conventionally it would be conceived, and stored on a computer, as a data set of 32*4,030 = 128,960 units of data.

The data contain relatively more information about the groups (19 variables for 99 units) than about the pupils (13 variables for 4,030 units). Arguably, the group-level variables are also more reliable because they refer to school or teacher records and are responses from adult professionals, whereas the responses of pupils are subject to test-performance variation, recall of family circumstances and arrangements, varying interpretations of the questionnaire items and so on. Moreover, the pupil-level variables, e.g., XROT, have a large-group level component of variation; groups vary a great deal in their composition (means, standard deviations, etc.) of these variables. Hence, not only the 19 group-level variables, but also, to some extent, the 13 pupil-level variables potentially explain group-level variation among the 99 groups, whereas only the 13 pupil-level variables explain some of the pupil-level variation in the outcome scores of the 4,030 pupils.

CHAPTER II: MODELS

Variance Component Models

The hierarchical structure of the data, with pupils nested within groups, requires a form of regression analysis that takes into account the two separate sources of variation in achievement. Separation of the variation attributable to pupils and to schools/classrooms is also of substantive interest, because the latter is a measure of the size of unexplained differences among schools/classrooms.

Goldstein (1986), Raudenbush and Bryk (1986) and Aitkin and Longford (1986) have established the relevance of variance component methods for

analyzing data with hierarchies. They address the previously mentioned problems with the use of ordinary regression methods when the assumption of independence of the observations is not satisfied.

Analytical Framework

Educational surveys involve hierarchically structured data--pupils within classrooms within schools within administrative units or regions. Every classroom (school, region) has its own idiosyncratic features that result from a complex of influences, including composition, teaching practices and management decisions. As a consequence, observations on students (e.g., their outcomes) are not statistically independent, not even after taking into account the available explanatory variables. This condition violates the assumption of independence for ordinary regression (OLS).

By comparison, variance component models are an extension of ordinary regression models that allow more flexible modelling of variation: within school or classroom and between schools or classrooms. Pupils are associated with (unexplained) variation, but this variation has a consistent within-classroom component that itself has a within-school component, etc. Schools vary, classrooms within schools vary and pupils within classrooms vary. Consider the regression model for data with two levels of hierarchy (pupils i within classrooms j):

$$y_{ij} = \alpha + \beta x_{ij} + \gamma z_{ij} + \epsilon_{ij} \qquad (1)$$

where α, β and γ are (unknown) regression parameters, x and z are explanatory variables, y is the outcome measure and the random term ϵ is assumed to be a

random sample from a normal distribution with a mean of zero and an unknown variance σ^2. Variation among the classrooms can be accommodated in the "simple" variance component model:

$$y_{ij} = \alpha + \beta x_{ij} + \gamma z_{ij} + a_j + \epsilon_{ij} \qquad (2)$$

where the a's form a random sample from a normal distribution with a mean of zero and an unknown variance τ^2 , and the a's and the ϵ's are mutually independent. The covariance of two pupils within a classroom is τ^2 (correlation $\tau^2/[\tau^2 + \sigma^2]$). If we knew the a's, we could use them to rank the classrooms. Model (2) has the form of analysis of variance (ANOVA) with distributional assumptions imposed on the a's. The advantages of this assumption are discussed by Dempster, Rubin and Tsutakawa (1981), who use the term "borrowing strength" in estimating the effects of small groups, and by Aitkin and Longford (1986).

In this model, each school has a uniform effect on the pupils within it. As this assumption may be unrealistic, a more flexible model is needed that allows not only the school means but also the school regression coefficients to vary, as some schools may be more "suitable" for pupils with certain backgrounds than others. This corresponds to variation in the within-school regressions of y on x and z. This situation can be suitably modelled as

$$y_{ij} = \alpha + \beta x_{ij} + \gamma z_{ij} + a_j + b_j x_{ij} + c_j z_{ij} + \epsilon_{ij} \qquad (3)$$

or

$$y_{ij} = \alpha + \beta x_{ij} + \gamma z_{ij} + a_j + b_j x_{ij} + e_{ij}. \qquad (4)$$

The classroom-level random effects (a_j, b_j) are assumed to be a random sample from a normal distribution with a mean of zero and an unknown variance $\Sigma^{(2)}$. Here $\Sigma^{(2)}$ involves three parameters: the variances of a and b and their covariance. Extensions to larger numbers of explanatory variables and to more complex hierarchies are described in the literature (e.g., Goldstein 1987; Longford 1987; Raudenbush and Bryk 1986).

The maximum likelihood estimation procedures for such models used in this paper are based on the Fisher scoring algorithm (Longford 1987) implemented in the software VARCL (Longford 1986). It provides estimates of regression parameters and (co-) variances, together with standard errors for them, and the value of the log-likelihood.

Variance Component Models Compared with OLS

Variance component methods involve the explicit modelling of student and group variation and afford flexibility in modelling the group variation, something that ordinary regression cannot do. The specification of a variance component model is necessarily more complex than is the case with ordinary regression. In standard situations, the analyst first declares the list of the regression variables involved in explaining the outcome for a typical group. Next the analyst declares a sublist of this list that contains the variables for which the within-group relationships are hypothesized to vary from group to group. The full list of variables, referred to as the "fixed part," is analogous to the list of the explanatory variables in ordinary regression. The sublist (random part) may contain only pupil-level variables, that is, variables that take on different values for students attending the

same class. Variables measured at the classroom level whose values are constant for all students in a classroom cannot be specified in the random part of the model, because within-group regression coefficients on group-level variables cannot be identified.

Variance component models involve two kinds of parameters. The fixed effects parameters refer to the regression relationship for the average group. Their interpretation is analogous to the regression parameters in ordinary regression. The random effects parameters are variances and covariances that describe the between-group variation in the regression relationship. Of prime interest are the sizes of the variances. Zero variance of a regression coefficient corresponds to a constant relationship across the groups. To obtain information about the variation, we require, in general, a substantially larger number of pupils and groups than we do for the regression parameters. We can therefore expect to find that a small random part, containing only a few variables, provides a sufficient description of the variation, whereas the fixed part may contain most of the available explanatory variables.

One important aspect of the separation of the two sources of variation is the ability to distinguish between pupil- and group-level variation. This aspect comes out very clearly in the following examples: it turns out that we have abundant group-level information, i.e., a good description of the between-group variation, but a much larger proportion of the student-level variation remains unexplained.

To fix ideas, we consider first a specific model:

$$y_{ij} = \Sigma_k x_{ij,k} \beta_k + d_j + \epsilon_{ij} \qquad (5)$$

where the indices $i = 1, \ldots, n_j$, $j = 1, \ldots, N_2$ and $k = 0, 1, \ldots, K$, represent the pupils, groups and variables, respectively. The β's are the regression parameters, and the d's and ϵ's are the group- and pupil-level random effects, assumed to be independent random samples from the normal distribution with zero means and variances σ^2 and τ^2. We will assume throughout that β_0 is the intercept, i.e., $x_{ij,0} = 1$. Analogously with the ordinary regression, we can define the R^2 as the proportion of variation, explained as

$$R^2 = 1 - (\sigma^2 + \tau^2)/(\sigma^2_{raw} + \tau^2_{raw}), \tag{6}$$

where the subscript "raw" refers to the variance estimates in the "empty" variance component model:

$$Y_{ij} = \mu + d_j + \epsilon_{ij}. \tag{7}$$

It is advantageous, however, to define two separate R^2s that refer to the two levels of the hierarchy for pupils and groups, respectively:

$$R_p^2 = (1 - \sigma^2)/\sigma^2_{raw} \tag{8}$$

$$R_g^2 = (1 - \tau^2)/\tau^2_{raw}. \tag{9}$$

CHAPTER III: SCHOOL EFFECTS ON MATHEMATICS LEARNING

Two questions that educators frequently ask are how much student achievement increases over the course of a year and whether schools affect growth in achievement differentially. In this section, we use the pretest (XROT) and student posttest (YROT) to address these questions. We also demonstrate, using simple examples from the data, the differences between ordinary regression, simple variance component analysis and variance component analysis using random coefficients. In the next section on the results of our analysis, we apply these techniques to the complete data set, using more complex models.

Model 1: Ordinary Regression (OLS)

In the present analysis, for a data set obtained by listwise deletion with respect to a set of variables considered below (a procedure that leaves 3,136 pupils in 88 schools), we have for the simple ordinary regression of posttest (YROT) on pretest (XROT), as per equation (1) with a single explanatory variable,

$$y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij} \tag{10}$$

and

$$\text{YROT} = 4.892 + .818\ \text{XROT}. \tag{11}$$
$$(.015)$$

In this model, identification of pupils within schools is completely ignored; instead, the pupils are assumed to be a randomly drawn sample from the population of all pupils in the given grade in the country. A pupil with a given pretest score XROT is expected to score 4.892 + .818XROT on the posttest. The standard errors for the regression estimates will be given throughout the paper in parentheses in the line below the regression parameters. For example, .015 above is the standard error for the regression coefficient on XROT, .818. The corresponding t-ratio is .818/.015 = 54.5.

The computation of R^2 follows:

$$\sigma^2_{raw} = 82.80$$
$$\sigma^2 = 42.56,$$

so that $R^2 = 1 - \sigma^2/\sigma^2_{raw} = 1 - (42.56/82.80) = .486$.

Model 2: (Simple) Variance Component Model (VCS)

To take into account the group-level variables, we choose a simple variance component model ("simple" in that it does not contain variable slopes):

$$Y_{ij} = \mu + d_j + \epsilon_{ij} \qquad (12)$$
$$\sigma^2_{raw} = 55.56$$
$$\tau^2_{raw} = 25.65.$$

The variation in posttest scores has a substantial group-level component. That is, the "total" variance is 81.21 (55.56 + 25.65), of which .316 (25.65/81.21), the variance component ratio, is attributed to group-level effects. The variance component regression model is given as:

$$\text{YROT} = 5.841 + .699\ \text{XROT} \qquad (13)$$
$$(.018)$$
$$\sigma^2 = 38.55$$
$$\tau^2 = 4.78,$$

so that we have $R^2 = 1 - (43.33/81.21) = .466$, and

$$R_p^2 = 1 - 38.55/55.56 = .306$$
$$R_g^2 = 1 - 4.78/25.65 = .814.$$

Thus, if we make allowances for the within-school correlation of the posttest scores, we obtain a prediction formula for the posttest score (YROT = 5.841 + .699XROT) that is substantially different from the OLS regression described in equation 11. Note, also, by how much the school-level variation has been reduced.

Table 3 presents the comparison between the simple OLS and simple variance component models. Clearly, the latter extension of the R^2 for variance components is more informative. The pretest score XROT is a powerful predictor of the posttest score YROT. However, whereas it explains more than 80% of the variation among the groups, it explains only 30% of the pupil-level variation. The school-level variation in the outcome scores reflects the pretest score to a great extent. Some of the remaining within-group variation may be explained by the other explanatory variables, but they are not likely to have as dominant an effect as the pretest score does.

The variation associated with the testing and scoring procedure, which could be demonstrated in an experiment with repeated administration of the test, use of alternate forms, etc., will remain as a component of the pupil-level variation. Thus, whereas the group-level variation can potentially be reduced to 0, the pupil-level variation has a component that cannot be explained by any explanatory variables. In ideal circumstances (and in our case, almost), we can explain completely why/how schools vary; the variance of schools in the later models is very small. We cannot, however, explain the pupil-level variation completely; there will always be an unexplainable within-pupil variation because of fluctuations in performance, distractions, guessing and so on. Since every pupil provides only one outcome score, the within-pupil and within-group variation cannot be separated.

The raw variance component ratio is .316, but with the model with the pretest score, the ratio drops to .110. If the pretest score is ignored, the groups appear to have substantial differences. At the same time, the schools appear to be much more similar (homogeneous) once we take account of the pretest scores, i.e., they are much more similar in the way they "convert" initial ability into outcome.

Table 3: Comparison of OLS and VCS Models of Grade 8 Mathematics Posttest Predicted from the Pretest, Thailand, 1981-82

Models	Method	
	OLS	VCS
Empty model		
σ^2_{raw}	82.80	55.56
τ^2_{raw}	-	25.65
Regression model		
Intercept	4.892	5.841
Coefficient	0.818	0.699
St. error coeff.	0.015	0.018
$\sigma 2$	42.56	38.55
τ^2	-	4.78
R^2	0.486	-
R_p^2	-	0.306
R_g^2	-	0.814

If a group-level explanatory variable were added to the regression model, it would result in a reduction of only the group-level variance, which has already been substantially reduced. Therefore there is less scope for important group-level explanatory variables than for pupil-level ones. Among the pupil-level variables there might be ones that explain a great deal of the remaining pupil-level variation.

Inclusion of a pupil-level variable in the regression model will cause a reduction in both the pupil- and group-level variances. The relative

sizes of the reductions of the two variances will depend on how the variation in the explanatory variable decomposes into between- and within-group variance. Hence, potentially the most important pupil-level explanatory variables are those with little between-group variation.

Model 3: Variable Slopes Model

The variance component model discussed above can be further generalized into a model that allows variable slopes on the pretest:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + d_{oj} + d_{1j}(x_{ij} - x) + \sigma_{ij}, \qquad (14)$$

where (d_{0j}, d_{1j}) form a random sample from a normal distribution with a mean of zero and an unknown variance, Σ_d; x is the sample mean for x; and ϵ's are a random sample from a normal distribution with a mean of zero and an unknown variance, σ^2. The maximum likelihood estimates for this model are:

$$\beta_0 = 5.832$$

$$\beta_1 = .687 \ (.019)$$

$$\sigma^2 = 38.367$$

$$\Sigma_d = \text{Var}\ (d_0, d_1) = \begin{matrix} 4.947 & \\ .0805 & .00416 \end{matrix} .$$

The software VARCL used for maximum likelihood estimation in variance component models estimates the square roots of the variances in Σ_d and produces standard errors for these estimates:

$\Sigma_{d,11}$ = 2.224 (.202)

$\Sigma_{d,22}$ = .0645 (.0338)

$\Sigma_{d,12}$ = .0805 (.0311).

Model 4: Comparison of the Models

Now we test Model 3 against Models 2 and 1. First, we compare Model 3 and Model 2. The value of the deviance (-2 log-likelihood)[6/] is 20,496.3. Using the conventional t-ratio, we conclude that the slope-variance $\Sigma_{d,22}$ is not significantly different from 0, so that we can adopt the simple variance component model.

More formally, we can use the likelihood ratio test to compare the two variance component models. The deviance for the simple Model 2 is 20,499.9, 3.6 times higher than in the case of the variable slopes Model 3. To determine the significance of this difference, it is necessary to determine the number of degrees of freedom from the "free" parameters. The simpler model is obtained from the latter model by constraining to zero the slope variance $\Sigma_{d,22}$ and the slope-by-intercept covariance $\Sigma_{d,12}$; these are the two additional free parameters that set the degrees of freedom equal to 2. Hence the statistic χ^2 has 2 degrees of freedom, and we can declare that we have found insufficient evidence for a variable slope of the posttest on the

---

[6/] This statistic is used to assess how well the model represents the data. For two models where one is a special case of the other, the differences of their deviances has a chi-square distribution, with the number of degrees of freedom equal to the difference in the number of free parameters in the two models.

pretest among the schools. That is, the schools are fairly uniform in their conversion of pretest scores into posttest scores.

Next we compare the simple variance component model (Model 2) with the ordinary regression model (Model 1). The differences among the schools, described by the variance τ^2 in the simple variance component model, are substantial and statistically significant; the formal likelihood ratio test for the hypothesis that $\tau_2 > 0$ is obtained by comparing the deviances of the ordinary regression and the simple variance component models. The ordinary regression deviance (-2 log-likelihood, which is not the same as the residual sum of squares) is equal to 20,662.6, 162.6 higher than the deviance for the simple variance component model (χ^2 with 1 degree of freedom). Therefore we reject the ordinary regression model in favor of the variance component model. Further, the t-ratio for τ is large.

Making inferences about relationships that vary from group to group is of substantive importance in studies of school effectiveness. Schools are expected to vary in their performance after accounting for differences in the initial ability of the pupils, but other more complex patterns of between-school variation may arise: schools may be relatively more successful in teaching children with certain background characteristics, and they may either exaggerate or reduce the differences among the pupils at enrollment.

The relationships among variables are intimately connected with variance heterogeneity. By way of illustration, we consider the variable slope model discussed above. The fitted variance of an observation is

$$38.367 + 4.947 + 2*(\text{XROT} - 8.912)*.08054 \qquad (15)$$
$$+ (\text{XROT} - 8.912)^2 *.00416.$$

It is a quadratic function of the pretest. The minimal variance occurs for XROT* = 8.912 - .0805/.0042 = -10.45 and is equal to 41.75. Only two pupils in the whole sample have scores lower than XROT*. Larger values of the explanatory variable XROT are associated with larger variance. For XROT = 9 (near the mean), the fitted variance is 43.33, and for XROT = 30 (near the sample maximum), the fitted variance is 48.56. It would appear that for low-ability pupils, the choice of school is slightly less important than for high-ability pupils. We have to bear in mind, however, that we are dealing with an observational study, not with an experiment, and in reality pupils, or their parents, do not have complete freedom of choice over the school. Thus a causal statement, or a prediction about a future manipulative procedure, can be made only under the condition that all the other circumstances in the educational system remain intact. This assumption is usually very unrealistic.

Summary

The comparison of the regression relationship (fixed effects) is instructive. We have

(i) Ordinary regression

$$YROT = 4.892 + .818*XROT$$
$$(.015)$$

(ii) Simple variance component model

$$YROT = 5.841 + .699*XROT$$
$$(.017)$$

(iii) Variable slopes

$$YROT = 5.832 + .687*XROT.$$
$$(.019)$$

The estimate of the regression coefficient on XROT in ordinary regression is substantially different from the estimates in the two variance component models. Ignoring the hierarchical structure of the data would lead to different conclusions, say, in predicting the posttest (YROT) from the pretest (XROT). In other words, whereas the OLS estimate could be interpreted to mean that each point on the pretest is worth .82 points on the posttest, the VCS estimate more accurately places this value at .69 points.

CHAPTER IV: PUPIL BACKGROUND AND SCHOOL/CLASSROOM EFFECTS ON LEARNING

Overview

In this section we use the complete data set to estimate the effects of student background and school/classroom variables on achievement in mathematics. The approach taken is often referred to as a "value-added" approach, since the purpose is to explain posttest achievement after the effects of prior learning (pretest achievement) have been taken into account. Our intent is to obtain the most parsimonious simple variance component model of grade eight mathematics learning in Thailand, given the data.

Because of missing data, we build the model conservatively, as follows. First, we start with the data set obtained by listwise deletion with respect to all 32 variables (including the outcome YROT and the pretest XROT),

fit a regression model to this data set, and apply a conservative criterion (to be specified below) to exclude variables from the obtained regression formula, so that we end up constructing a restricted set of explanatory variables. We apply listwise deletion to this restricted set of variables, a process that leads to a larger sample of pupils and schools. For this new data set, we again fit the regression model, simplify the regression formula, if possible, and continue on until no further reduction of the set of variables and extension of the data set obtained by listwise deletion are possible.

Usually it cannot be assumed that the unavailable data are missing at random, i.e., the distribution of a variable among the pupils from whom we obtain valid responses is similar to the distribution among the pupils whose responses are not available (missing). In educational surveys, typically higher ability pupils, those with higher social status, etc., tend to have higher response rates, the implication being bias in the estimates of certain population means, as well as in the regression coefficients obtained from simple regression. Missingness at random is an unnecessarily stringent criterion for ensuring that the omission of the subjects with missing data has no effect on the results of a regression analysis. It is sufficient to have conditional randomness, given the explanatory variables. It means that for any combination of explanatory variables, the distribution of the outcome among the pupils in the sample is identical to that for those excluded from the sample by the listwise deletion procedure. Intuitively, such an assumption becomes less stringent the more explanatory (conditioning) variables are used. On the other hand, a larger set of explanatory variables

implies a larger proportion of subjects whose data are not used in the analysis.

An indication of the extent to which the criterion of conditional randomness is relevant can be deduced from comparisons of model fits for two different samples: the maximal sample obtained by listwise deletion with respect to the set of explanatory variables used in the considered model, and the sample obtained by listwise deletion with respect to a more extensive, or complete, set of explanatory variables. In a few such comparisons, reported below, we find close agreement in several pairs of such analyses.

Multiple Regression Models

The response rate for the 13 pupil-level variables is between 93-100%. There is no obvious pattern of missingness among the pupils; complete pupil-level records are available for 3,466 individuals (86%). The group-level data are available for between 78-99 schools, but only 60 schools have complete records, and within these schools, only 2,076 pupils also have complete pupil-level data (51.5%). We begin by fitting the simple variance component models (VCS), i.e., models involving no variable slopes, to the data set.

First model: Regression with all variables. Listwise deletion with respect to all 32 available variables results in a data set containing 2,076 pupils in 60 schools. The ordinary regression fit (OLS) of the posttest on the pretest is

$$\text{YROT} = 4.882 + .817*\text{XROT}, \qquad \sigma^2 = 42.20,$$
$$(.017)$$

which is in close agreement with the OLS fit reported above for the larger data set (3,136 pupils in 88 schools). The corresponding simple variance component model fit is:

$$\text{YROT} = 5.670 + .720*\text{XROT}$$
$$(.020)$$
$$\sigma^2 = 38.79$$
$$\tau^2 = 4.02.$$

Compared to the larger data set, equation 13, we find some discrepancies: the fitted regression slope for the smaller data set is higher (.720 versus .699) and the group-level variance is smaller (4.02 versus 4.78). The variation of the slope on XROT is not significant in either sample, but it is two-and-a-half times as great in the larger data set (.00416) than in the smaller one (.00166). It appears that the 28 schools added to the data are more likely to have lower regression slopes and contain proportionately more schools at the extremes (very "good" or very "bad"), because the larger sample has larger group-level variance, τ^2. We emphasize that all these differences may arise purely by chance, rather than as a result of non-random missingness of the data, but they can have a substantial effect on the inferences drawn.

The OLS and VCS model estimates for the 2,076/60 data using all the explanatory variables are given in Table 4. The dominant explanatory power of the pretest score XROT is obvious, as evidenced not only by the t-ratio for its regression coefficient (32.38 for OLS and 30.80 for VCS), but also by the comparison of the variance component estimates across models. The raw variance component estimates are:

$$\sigma^2_{raw} = 57.30$$
$$\tau^2_{raw} = 28.83.$$

Table 4: OLS and VCS Model Estimates for 2,076 Students and 60 Classrooms/Schools Using All 31 Explanatory Variables, Thailand, 1981-82

Variable	OLS		VCS	
	Estimate	St. Error	Estimate	St. Error
Student Level				
GRAND MEAN	18.603	-	19.717	-
XROT	.680	.021	.647	.021
XAGE	-.080	.016	-.077	.016
XSEX	.732	.301	.969	.319
YFOCCI	.174	.431	.033	.434
	-.631	.462	-.646	.460
	-.178	.541	-.239	.542
YMEDUC	.021	.327	-.039	.325
	-.129	.562	-.157	.556
	-.686	.661	-.899	.663
HCALC	-.120	.310	-.217	.309
YHLANG	.203	.315	.012	.341
YMOREED	1.087	.546	1.074	.541
	1.570	.545	1.537	.541
	1.638	.593	1.610	.589
YPARENC	.225	.137	.249	.136
YPERCEV	-.980	.160	-1.020	.161
YFUTURE	.574	.168	.526	.167
YDESIRE	.277	.236	.228	.233
Group Level				
spci81	.061	.042	.073	.060
senrolt	.422	.263	.417	.386
sstream	-.426	.358	-.500	.512
sdaysyr	-.006	.020	-.010	.029
sputear	-.152	.051	-.170	.075
squalmt	1.023	.342	1.029	.494
tedmath	-.035	.037	-.044	.053
tsex	-.580	.336	-.619	.481
tage	.009	.032	-.001	.046
texptch	.014	.043	.038	.064
tnstuds	.035	.018	.039	.025
tmthsub	1.725	.432	1.941	.628
txtbook	1.602	.338	1.650	.490

(continued)

Variable	OLS Estimate	OLS St. Error	VCS Estimate	VCS St. Error
cefeed	.148	.203	.209	.290
tworkbk	-1.104	.218	-1.124	.314
tvismat	.380	.331	.461	.480
tadminl	-.003	.004	-.003	.006
torderl	-.037	.012	-.039	.016
tseatl	.011	.005	.011	.007
Variance	38.031	6.167	-	-
Pupil-level variance	-		36.809	-
Pupil-level sigma	-		6.067	-
Group-level variance	-		1.317	-
Group-level sigma	-		1.148	0.192
Deviance	-		13424.947	-

The pretest score XROT on its own leads to a reduction of these variances to 38.79 (R_p^2 = 32%) and 4.02 (R_g^2 = 86%). However, the other 30 variables reduce the pupil-level variance only marginally to 36.8 (R_p^2 = 36%). The group-level variance is almost saturated—1.32 (R_g^2 = 95.5%). It appears that we have abundant information about the groups, but we are less successful with an explanation, or suitable description, of the pupil-level variation.

The relatively large number of group-level variables raises a concern about multicollinearity, i.e., competing alternative descriptions of the data. To deal with this problem we apply a conservative criterion for the exclusion of explanatory variables from our models. We regard a variable as not "important" for the fixed part of the VCS model if the t-ratio of its regression coefficient is smaller than 0.9 at the first stage of model reduction and 1.0 thereafter. In the first round of simplifying the model, we use the 0.9 criterion to exclude two pupil-level social class variables (calculator in the home [YHCALC] and use of the language of instruction in the

home [YHLANG]) and six group-level variables: four indicators of resource inputs (number of days in the school year [sdaysyr], teacher's postsecondary mathematics education [tedmathl], teacher's age [tage], and teaching experience [texptchl]) and two teaching process variables (frequent use of individual feedback [cefeed] and time spent in routine administration [tadminl]) from the full list of 31 variables.

Second model. Next we estimate both the OLS and VCS models using this shorter list of 23 variables. The results are shown in Table 5. Exclusion of the eight variables (eight degrees of freedom) has virtually no effect on the retained regression parameters and their standard errors (compare Tables 4 and 5); the exception is an indicator of instructional materials (use of commercial visual materials [tvismat]), which now fails to meet the inclusion criterion. The increase in the variance components is only marginal, in particular for the group-level variance. The difference in deviances is 3.3 (χ^2_8).

Again we obtain the largest data set obtainable by listwise deletion with respect to the retained variables; this procedure yields data for 2,804 pupils in 80 schools. We then compute the variance component analysis for this data set; the results are given in Table 6. We see that the regression coefficients for the pupil-level variables are stable across the data sets (as compared with Tables 4 and 5), but the discrepancies for the group-level variables are substantial. There are two separate, but possibly complementary, explanations for these discrepancies: multicollinearity and non-random missingness of data. Multicollinearity would cause the regression estimates to be sensitive to changes in the data, in our case to the inclusion

Table 5: OLS and VCS Model Estimates for 2,076 Students and 60 Classrooms/Schools Using 23 Explanatory Variables, Thailand, 1981-82

Variable	OLS		VCS	
	Estimate	St. Error	Estimate	St. Error
Student Level				
GRAND MEAN	18.118	-	18.370	-
XROT	.685	.020	.650	.021
XAGE	-.080	.016	-.076	.016
XSEX	.723	.299	.958	.318
YFOCCI	.118	.426	.033	.432
	-.621	.457	-.651	.457
	-.139	.538	-.212	.541
YMEDUC	.037	.326	-.028	.325
	-.068	.559	-.115	.555
	-.604	.656	-.855	.660
YMOREED	1.115	.545	1.083	.540
	1.568	.543	1.521	.540
	1.666	.591	1.609	.589
YPARENC	.238	.137	.255	.135
YPERCEV	-.970	.160	-1.010	.161
YFUTURE	.570	.168	.526	.167
YDESIRE	.287	.235	.234	.233
Group Level				
spci81	.050	.038	.058	.056
senrolt	.509	.251	.540	.373
sstream	-.441	.324	-.503	.472
sputear	-.178	.046	-.198	.068
squalmt	1.062	.327	1.090	.480
tsex	-.518	.314	-.536	.460
tnstuds	.036	.017	.038	.025
tmthsub	1.802	.409	2.094	.604
txtbk	1.649	.315	1.673	.463
tworkbk	-1.028	.204	-1.039	.300
tvismat	.368	.322	.393	.473
torderl	-.040	.010	-.043	.014
tseatl	.010	.005	.011	.007
Variance	38.108	6.173	-	-
Pupil-level variance	-		36.855	-
Pupil-level sigma	-		6.071	-
Group-level variance	-		1.351	-
Group-level sigma	-		1.162	.191
Deviance	-		13428.295	-

of over 700 new observations. As an alternative, the discrepancies could arise as a result of the non-random missingness in our data, i.e., if the two data sets have genuinely different regression characteristics. A suitable indication, although not a fool-proof check, for the latter possibility is obtained by fitting the models with identical specifications for the different "working" data sets. We have fitted the reduced second model (Table 5) to the larger data set (Table 6), and although we obtained different values for the group-level regression coefficients, it turns out that the reduced list of variables also provides an adequate description for the data (as judged by the likelihood ratio criterion). The pupil-level regression coefficients differ only marginally.

We conclude, therefore, that multicollinearity is the more likely cause of the discrepancies in the estimates: we have too many group-level variables, so that the parameter estimates are subject to large fluctuations when small changes are made in the data. The explanatory variables provide sufficient conditioning for the outcome data to be missing at random, given the available explanatory variables.

In keeping with According to our exclusion criterion (t ratio < 1), we now delete from the fixed part of the model six group-level variables. Four are conventional material and non-material input variables (district level per capita income [spci81], teacher gender [tsex], class size [tnstuds], and use of commercial visual materials [tvismat]) and two are organization and process variables (student time doing seatwork [tseat1] and ability grouping [sstream]).

Table 6: OLS and VCS Model Estimates for 2,804 Students and 80 Classrooms/Schools Using 23 Explanatory Variables, Thailand, 1981-82

Variable	OLS		VCS	
	Estimate	St. Error	Estimate	St. Error
Student Level				
GRAND MEAN	17.659	-	17.314	-
XROT	.699	.017	.634	.019
XAGE	-.079	.014	-.073	.014
XSEX	.746	.251	1.103	.271
YFOCCI	.197	.363	.101	.367
	-.403	.389	-.458	.386
	.089	.458	.085	.458
YMEDUC	.306	.279	.293	.276
	.088	.465	.142	.458
	-.018	.567	-.309	.566
YMOREED	.861	.476	.786	.467
	1.086	.475	1.015	.468
	1.617	.519	1.542	.512
YPARENC	.388	.118	.375	.116
YPERCEV	-1.083	.137	-1.131	.136
YFUTURE	.576	.142	.533	.141
YDESIRE	.493	.201	.439	.198
Group Level				
spci81	-.029	.033	-.025	.057
senrolt	.437	.187	.481	.331
sstream	-.417	.275	-.422	.473
sputear	-.095	.032	-.110	.058
squalmt	.698	.246	.784	.429
tsex	-.038	.266	.014	.463
tnstuds	.012	.014	.020	.023
tmthsub	1.836	.344	2.398	.593
txtbk	.948	.266	.978	.461
tworkbk	-0.500	.167	-.499	.291
tvismat	.353	.269	.363	.468
torderl	-.024	.008	-.027	.013
tseatl	.005	.004	.006	.006
Variance	37.949	6.160	-	-
Pupil-level variance	-		35.868	-
Pupil-level sigma	-		5.989	-
Group-level variance	-		2.285	-
Group-level sigma	-		1.512	0.174
Deviance	-		18088.395	-

Third model. As before, we estimate this model with both the smaller and larger data sets. The estimates from the OLS and VCS models using the former reduced list of variables are given in Table 7; the same schools and pupils are involved as for Table 6. For the latter, larger data set of 3,025 students in 86 schools, we fit the reduced model (17 variables) and present the results in Table 8. Again, the difference in deviances (3.5, χ_6^2) is small. The effects of non-random missingness can be checked by comparing the estimates in Tables 7 and 8. Applying our exclusion criterion to the variables in Model 3, we find that no further reduction of the list of explanatory variables is possible.

Note that, because of the relatively small number of schools, the appropriate conclusion about the 14 group-level variables we deleted is that "we found insufficient evidence" of a systematic effect of these variables, rather than "our analysis disproves their effects." Further, a different modelling scheme could lead to a different "minimal" set of important explanatory variables. Because of collinearity, there may be a set of alternative regression formulae that give a model fit that is not substantially inferior to the one given in Table 8 in terms of the deviances.

A summary of the results of these analyses is provided in Table 9. In all the models, student background characteristics are important determinants of mathematics learning over time. School-level resources also appear to have an important impact on achievement, with students in the larger schools learning more than students in the smaller schools and students in schools with a higher percentage of teachers qualified to teach mathematics learning more than students in schools with a lower percentage of qualified teachers; however, students in the schools with a higher student/teacher ratio also learned more.

Table 7: OLS and VCS Model Estimates for 2,804 Students and 80 Classrooms/Schools Using 17 Explanatory Variables, Thailand, 1981-82

Variable	OLS		VCS	
	Estimate	St. Error	Estimate	St. Error
Student Level				
GRAND MEAN	17.321	-	17.694	-
XROT	.704	.017	.635	.018
XAGE	-.077	.014	-.073	.014
XSEX	.676	.247	1.086	.270
YFOCCI	.181	.357	.085	.365
	-.419	.387	-.465	.385
	.105	.455	.082	.457
YMEDUC	.293	.280	.288	.276
	.112	.465	.154	.458
	.014	.563	-.297	.564
YMOREED	.869	.476	.786	.467
	1.128	.476	1.027	.468
	1.666	.520	1.560	.512
YPARENC	.393	.117	.377	.116
YPERCEV	-1.076	.137	-1.130	.136
YFUTURE	.592	.142	.537	.141
YDESIRE	.477	.201	.431	.197
Group Level				
senrolt	.285	.164	.367	.289
sputear	-.074	.030	-.094	.054
squalmt	.808	.239	.880	.427
tmthsub	1.950	.329	2.562	.576
txtbook	.948	.259	.946	.458
tworkbk	-.433	.160	-.402	.284
torderl	-.022	.006	-.024	.010
Variance	38.065	6.170	-	-
Pupil-level variance	-		35.871	-
Pupil-level sigma	-		5.989	-
Group-level variance	-		2.429	-
Group-level sigma	-		1.558	0.176
Deviance	-		18091.983	-

Table 8: OLS and VCS Model Estimates for 3,025 Students and 86 Classrooms/Schools Using 1/ Explanatory Variables, Thailand, 1981-82

Variable	OLS		VCS	
	Estimate	St. Error	Estimate	St. Error
Student Level				
GRAND MEAN	17.238	-	17.536	-
XROT	.695	.017	.629	.018
XAGE	-.075	.014	-.071	.014
XSEX	.658	.238	1.053	.260
YFOCCI	.152	.343	.074	.351
	-.415	.373	-.435	.373
	.115	.443	.123	.446
YMEDUC	.371	.269	.343	.265
	.056	.449	.073	.442
	.066	.554	-.259	.555
YMOREED	.854	.461	.755	.453
	1.195	.459	1.064	.452
	1.703	.500	1.532	.494
YPARENC	.361	.113	.347	.112
YPERCEV	-1.140	.132	-1.191	.132
YFUTURE	.614	.137	.543	.136
YDESIRE	.484	.194	.459	.190
Group Level				
senrolt	.271	.160	.350	.279
sputear	-.076	.029	-.094	.052
squalmt	.847	.232	.903	.410
tmthsub	1.968	.327	2.546	.566
txtbk	1.047	.250	1.071	.437
tworkbk	-.434	.157	-.417	.275
torderl	-.023	.006	-.025	.010
Variance	38.271	6.186	-	-
Pupil-level variance	-		36.138	-
Pupil-level sigma	-		6.012	-
Group-level variance	-		2.353	-
Group-level sigma	-		1.534	.169
Deviance	-		19537.962	-

Classroom variables also affect achievement. Students in non-remedial classes learned more than students in remedial classes; students in classes where the teacher used textbooks more often learned more than students in classes in which textbooks were not used. On the other hand, workbooks and teacher time spent maintaining order were negatively related to learning.

Table 9: Summary of Tables

	Tables				
	4	5	6	7	8
OLS variance	38.03	38.11	37.95	38.07	38.27
St. error	6.17	6.17	6.16	6.17	6.19
VCS pupil-level variance	36.81	36.96	35.87	35.87	36.14
Sigma	6.07	6.08	5.99	5.99	6.01
VCS group-level variance					
For G. mean	1.32	1.35	2.29	2.43	2.35
Sigma	1.15	1.16	1.51	1.56	1.53
St. error for sigma	0.19	0.19	0.17	0.17	0.17
Sample size					
Pupils	2,076	2,076	2,804	2,804	3,025
Groups	60	60	80	80	86

Several researchers have considered the contextual effects in educational studies involving multi-level data (see Raudenbush and Bryk 1986). In our case, contextual analysis would involve using within-school means of pupil-level variables as school-level variables. However, as was pointed out earlier, we have abundant school-level information (14 school-level variables

for 99 schools), and contextual analysis would only aggravate further the high level of confounding of the school-level variables. Contextual variables are more relevant in studies where the aim is to produce, or at least consider, a ranking of schools. The ranking may depend crucially on the explanatory variables used and can often be affected by even the inclusion of variables with statistically insignificant regression coefficients. This point highlights the need to select models based on educational theory rather than on purely statistical criteria that contain a great deal of arbitrariness.

Modelling of Group-Level Variation (Random Slopes and Random Differences)

Simultaneously with reducing the fixed (regression) part of the variance component model for our data, we also need to explore extensions of the random part to obtain a better description of the group-level variation than the one offered by the group-level variance. We concentrate first on a reduction of the fixed part to a shorter list of explanatory variables because: (i) the school-level variation is rather small and (ii) in the models with complex descriptions of variation, the estimates of fixed effects and their standard errors differ very little from those obtained so far (Table 8).

In the variance component models fitted so far (Tables 4-8), the within-group regressions are assumed to be constant across groups, with the exception of the intercept (position), which has a fitted variance of 2.35. More generally, the regression coefficients with respect to any of the pupil-level variables may be allowed to vary across the groups. These variables, selected from the variables included in the fixed part, form the random part of the model. The group-level variables are not considered for

the random part, because within-group regressions with respect to such variables cannot be identified.

Variance component models closely resemble the models for the analysis of covariance. The simple variance component models correspond to ANOCOVA models, with no interactions of covariates with the grouping factor. The (complex) variance component models with variable within-group regressions (slopes and/or differences) correspond to ANCOVA models with group x covariate interactions. The difference between the variance component and ANOCOVA models is in their emphasis on the description of variation as opposed to differences among the groups and in the assumption of the normality of the group effects in the former. The model specification in both models is analogous:

. a, list of covariates (fixed part),

b, sublist of covariates that have interactions with the grouping factor (random part).

We now turn to modelling the random part. For a continuous variable included in the random part, the within-group regression slopes with respect to this variable are assumed to vary randomly (and to be distributed normally) with an unknown variance. For a categorical variable included in the random part, the within-group (adjusted) differences among the categories are normally distributed. We can consider the "stereotypical" group, for which the regression is given by the fixed part model (the average regression), with the regressions for the groups varying around this average regression. The deviations of the regression coefficients form a random sample (i.i.d.) from a multivariate normal distribution. The components of the vector of deviations (for a group) cannot be assumed to be independent; thus, their

covariance structure has to be considered. However, the variances of these deviations (or random effects) are the main interest.

Data with only a moderate number of groups and with limited numbers of subjects within groups (classroom sizes), as is the case in this analysis, contain only limited information about variation, comparable to the limited information about interactions in models of analysis of covariance. Usually, information about the covariance structure is even scarcer. Therefore, if many variances are included in the random part (and estimated as free parameters), we can expect high correlations among the estimates — large estimated variances with large standard errors. Moreover, the number of covariances to be estimated grows rapidly with the number of variances, and many of the estimated correlations corresponding to these covariances are then close to +1 or −1. The variance matrix with these variances and covariances is not of full rank, and the random effects are linearly dependent. Therefore it is important to adhere to the principle of parsimony and seek the simplest adequate description for group-level variation. In selecting the covariances to be estimated, we use the guidelines set by Goldstein (1987) and Longford (1987).

Although selection of a model for the random part involves only pupil-level variables (inclusion/exclusion), it is more complex than the selection for the fixed part because constraints can also be imposed on the covariances. The most general variance component model would involve 17 variances (the number of regression parameters in Table 8) and 17*16/2 = 136 covariances. Fitting such a model is clearly not a realistic proposition. Thus, model selection has to proceed by building up the random part from simpler to more complex models. The models fitted are all invariant with

respect to the choice of the location of the explanatory variables. In the computations, all the variables are centered around the overall mean, and the estimated variance matrix refers to this "centered" parametrization. However, the variance matrix for a different parametrization is easy to calculate by a quadratic transformation.

In selecting the model for the random part, we proceed according to the following stages. For all the models we use the same fixed part as in Table 8. The estimates and standard errors for the regression parameters differ very slightly from those in Table 8 for all these models. This fact justifies post hoc our approach of first settling the fixed part and then modelling the random parts. First we fit models with one pupil-level variable in the random part. Using the likelihood ratio test to compare the fitted model to the model with the simple random part (Table 8), we select the following variables: pretest score (XROT); age (XAGE); motivation (YDESIRE); and educational expectation (YMOREED).

The first three variables are ordinal and are associated with one variance each. The likelihood ratio (the difference of the deviances) for each of the three corresponding models is larger than 3. This criterion is intentionally very conservative, since we prefer to err on the side of inclusion. Two parameters are involved — a variance (slope-variance) and a covariance (slope-by-intercept covariance) — but they are not free parameters, since they have to satisfy the condition of positive definiteness. The distribution of the difference of the deviances is χ^2_2 only if the correlation corresponding to the covariance is not equal to +1 or -1. The problem of negative variances is resolved by estimating the square roots of the variances (sigmas). In the actual computational algorithm, negative

sigmas do not arise, and the estimated variance matrix is always non-negative definite.

Next we fit the VC model with these four variables in the random part and simplify the random part by excluding variables and setting certain covariances to 0. The variance associated with the variable XAGE is very small (.00095), and its square root has a low t-ratio (.75), so that it can be constrained to 0 (excluded). The implication is a constraint on all the covariances involving XAGE, which are also set to 0. The three remaining variables and the intercept are represented by a 6x6 variance matrix: 6 variances and 15 covariances, almost as many parameters as are in the fixed part. The fitted variance matrix is:

Intercept		2.581					
XROT		.0143	.00558				
YMOREED	Cat.2	.191	.0388	.812			
	Cat.3	.519	.0439	.0621	1.032		
	Cat.4	.384	.0354	-.0241	.261	1.032	
YDESIRE		.0863	-.0127	-.307	-.303	-.346	.677

The decrement in deviance as compared with the VCS model (Table 8) is only 13, a result that hardly warrants the addition of these 21 parameters in the model.

The software used provides standard errors for the square roots of the variances (sigmas and diagonal elements of the matrix) and for the covariances. The sigmas and their standard errors are:

	Intercept	XROT	YMOREED			YDESIRE
			cat. 2	cat. 3	cat. 4	
Sigma	1.607	.0747	.901	1.175	1.016	.828
St. error	.176	.0261	.429	.451	.640	.295

The standard errors for the covariances involving XROT and categories of YMOREED (rows 3-5 in column 2) are between .059 - .063 and for those involving YDESIRE and YMOREED (columns 3-5 in row 6) are .56 - .62. Since each of these covariances has a small t-ratio, they are constrained to 0 in the next model. The following estimated variance matrix is obtained (the sigmas and their standard errors are given to the right of the variance matrix):

Variable	Matrix						Sigma	St. Error
Intercept	2.237						1.496	.173
XROT	.0141	.00343					.0586	.0317
YMOREED Cat. 2	.199	0	.0230				.152	.639
Cat. 3	.601	0	.0791	1.490			1.221	.439
Cat. 4	.443	0	.003	.392	.826		.989	.753
YDESIRE	.119	-.0178	0	0	0	.746	.864	.276

Exclusion of these six covariances leads to an increase in the deviance of only 1.8. The variance associated with the second category of YMOREED falls substantially, and it can also be constrained to 0, together with the three covariances in the same row and column of the variance matrix. Constraining these four parameters causes an increase in the deviance of only .2. The reestimated variance matrix is:

Variables	Matrix						Sigma	St. Error
Intercept	2.415						1.554	.162
XROT	.0455	.00390					.0625	.0313
YMOREED Cat. 2	0	0	0				0	0
Cat. 3	1.136	0	0	1.788			1.337	.341
Cat. 4	.740	0	0	1.157	1.424		1.193	.514
YDESIRE	.304	-.0436	0	0	0	.830	.911	.260

The rank of this matrix is 4 (the two variance matrices given above are also singular). Thus it appears that another variance parameter can be constrained to 0. However, the t-ratio for each of the sigmas is high, and only a complex linear reparametrization of the variables included in the random part would enable further simplication of the model.

The variance matrix obtained provides a description of group-level variation in terms of 11 parameters, 5 variances and 6 covariances. However, the difference between the variances in this model and the corresponding VCS model is only 11 (for 10 parameters). That result provides further evidence

of overparametrization or collinearity in the random part. However, any attempt to define a suitable model with fewer parameters would necessarily involve some unnaturally defined variables, which would make interpretation of the model very difficult. We interpret these estimates as discussed below.

The variation in the slope on XROT provides evidence of an unequal "conversion" of ability at the beginning of the year into ability at the end of the year. Such a conclusion is appropriate only subject to the caveats discussed in the summary chapter. The slope on XROT is shallower in some schools, where the initial differences in XROT tend to be associated with smaller differences in YROT than in schools where the slopes are steeper.

The regression slope for YDESIRE is about .5, which is the regression slope for the "stereotypical" school, where every feature is "average." The variation associated with this regression slope has a standard deviation of .9; that is, there is a large (predicted) proportion of schools where the slope on YDESIRE is very small or even negative. The correlation of the within-group slopes on XROT and YDESIRE is -.77: lower "effects" of motivation to succeed are associated with schools where the initial differences become exaggerated by the end of the year.

The variances associated with categories 3 and 4 of YMOREED (expectations to complete five or more years of schooling) represent the variation of the adjusted differences between categories 3 and 1 (expectation to complete fewer than two more years of education) and 4 and 1, respectively. While the fitted difference between categories 2 (two to four more years) and 1 is about .8 and constant for all the schools, the average within-school difference between categories 3 and 1 is 1.1, with a variance of 1.8. Therefore this difference is negative in several schools. The situation with

the categories 4 and 1 contrast is similar, although the number of schools with the reversed sign of the difference is much smaller. The correlation of the random effects associated with categories 3 and 4 is .725; a high 3-1 contrast is associated with a high 4-1 contrast; but the fitted variance for the contrast 4-3 is 1.79 + 1.42 - 2*1.16 = .89, whereas the average difference is 1.58 - 1.08 = .50. Hence there are schools where the pupils with YMOREED = 3 have lower adjusted scores on YROT than where YMOREED = 4, although on average the fourth category is .5 points ahead.

The estimates of the regression parameters differ only marginally for the different specifications of the random part. This result justifies, post hoc, our approach of modelling first the regression part of the model and then the random part. The regression estimates for the last model considered are given in Table 10.

Conditional Expectations of the Random Effects

In the fixed-effects ANOVA or ANOCOVA, estimates of the effects associated with the groups are obtained. In variance component models, these effects are represented by random variables. Conditional upon the adopted model, the expectations of the (random) group-effects can be considered as the group-level residuals, or as "estimates" of the group-effects. These conditional expectations have to be inspected as to whether they conform with the assumptions of normality. This inspection involves a check for skewness and kurtosis (not carried out here, but visual inspection indicates no problems) and a check for outlying values of the effects. The latter check is obviously also of substantive importance because it would be useful to detect schools with exceptionally high or low performance, where the categories of

Table 10: Fixed-effect Estimates for the Final Model with Random Effects for 3,025 Students and 86 Classrooms/Schools Using 18 Explanatory Variables, Thailand, 1981-82

Variable	VCS Estimate	VCS St. Error
Student Level		
GRAND MEAN	16.642	-
XROT	.617	.020
XAGE	-.070	.014
XSEX	1.143	.260
YFOCCI	.101	.352
	-.488	.374
	.198	.446
YMEDUC	.347	.268
	.062	.446
	-.491	.560
YMOREED	.816	.453
	1.117	.476
	1.618	.514
YPARENC	.358	.112
YPERCEV	-1.178	.133
YFUTURE	.526	.137
YDESIRE	.480	.217
Group Level		
senrolt	.300	.265
sputear	-.063	.048
squalmt	.781	.380
tmthsub	2.632	.582
txtbook	0.949	.431
tworkbk	-.372	.270
torderl	-.035	.012
tseatl	.007	.006
Variance	-	-
Pupil-level variance	35.259	-
Pupil-level sigma	5.938	-
Group-level variance	See matrix in the text	
Group-level sigma		
Deviance	19,064.902	-
Number of iterations	8	

YMOREED have substantially different differences than do average schools, in which the outcomes are more/less influenced by the initial score XROT.

The complex nature of the variation, involving three variables, coupled with the number of groups, makes it infeasible to discuss the deviations of the group-level regressions from the average regression. In fact, the main motivation for using variance component analysis has been to obtain a global description of variation, without reference to individual groups. The added advantage is that owing to the shrinkage property of the conditional expections, extreme results attributable to unreliability for some of the schools with small numbers of students are avoided. The conditional expectations are a mixture of the pooled ordinary least squares solution and the within-group regression; the weight depends on the amount of information contained in the data from the group. Conditional expectations are obtained even for schools where the number of pupils in the data is smaller than the number of regression parameters. Because of this shrinkage, we cannot pinpoint all the schools where, say, the difference between categories 3 and 1 has a negative sign. For several schools, the conditional means indicate a small difference among the categories; some of these may be negative, others positive and larger than the conditional expectation. Accordingly, we should downscale our notion of what is an exceptionally large deviation; for example, a 1.5 multiple of the standard deviation (sigma) should be regarded as exceptional.

We conclude with an example of an exceptional school. All the random-effects components of school 22 (42 pupils in the data) are positive. Its deviation from the average regression formula is

$$1.517 + .100\ \text{XROT} + .102\ \text{YDESIRE} + 1.008\ YM_3 + .842\ YM_4,$$

where YM_3 (and YM_4) are equal to 1 if the pupil is in category 3 (4) and 0 otherwise. This outcome indicates that school 22 is characterized by high performance, with the differences in initial ability tending to get exaggerated. That is, pupils with high motivation and high expectations are at an advantage. For sample mean values of XROT and YDESIRE, this formula becomes

$$2.959 + 1.008\ YM_3 + .842\ YM_4,$$

which reflects the high "performance" of the school much more clearly. The variances quoted above refer to the regression using centered versions of all the variables (XROT - $\overline{\text{XROT}}$, YDESIRE - $\overline{\text{YDESIRE}}$, $YM_3 - \overline{YM_3}$, $YM_4 - \overline{YM_4}$). In the transformation from one parametrization to the other, only the intercept-variance is affected.

CHAPTER V: DISCUSSION

At the outset of this paper, we posed a series of questions: (i) do schools affect student learning differentially? (ii) what part of this variation is attributable to between school characteristics versus between student characteristics? (iii) what characteristics of teachers and schools enhance student achievement, independent of student background? (iv) are

these effects uniform across students? (v) what is the comparative effectiveness of alternative inputs? and (vi) how do estimates obtained from simple OLS methods compare with estimates obtained from multi-level methods? During the analysis, a sixth question arose: are there alternative regression models that predict student achievement equally well as the model developed herein? In this section, we review our findings and present some caveats about their interpretation.

Summary

School effects. The first analysis in this paper examined the extent to which schools differed in their ability to transform pretest scores into posttest scores. We found that the schools in this sample from Thailand were equally effective in converting pretest into posttest scores and that there were essentially no variable slopes in this respect. That is, the results from the simple variance component model did not differ significantly from those obtained from the variance component model that included variable slopes.

Contribution of school versus individual characteristics. In our second analysis, we examined group and individual effects on total variance. Group-level effects contributed 32% of the variance, while individual-level effects contributed 68% of the variance in posttest scores, after controlling for the pretest scores. We were able to explain most of the group-level variation but were less successful in explaining individual variation.

Effective teacher and school characteristics. The results from our final analysis indicate that some teacher and school characteristics are positively associated with student learning in Thailand:

- The percentage of teachers in the school that are qualified to teach mathematics
- an enriched mathematics curriculum and
- the frequent use of textbooks by teachers.

At the same time, some teaching practices are negatively related to learning:

- the frequent use of workbooks, and
- time spent maintaining order in the classroom.

The positive results are not surprising. Teachers who know the subject matter being taught, a curriculum that covers the domain, and textbooks that provide a structured presentation of the material all should have positive effects on achievement. The negative results are also unsurprising. Teachers who spend a great deal of time maintaining classroom order will have less time available for teaching; therefore, less learning takes place. Similarly, frequent use of workbooks may detract from effective teaching, answering questions and so forth.

Uniformity of effects. In this sample, we found that the schools did not have uniform effects on all students. In particular, the effects differed according to the level of students' expectations about further education. Some schools/classrooms were more effective for students with low expectations, some were more effective for students with high expectations, while others were equally effective (or ineffective) for all types of students. Interestingly enough, we found little evidence that schools were differentially effective for students on the basis of gender, age, parental occupation or several other student attitudes.

Comparative effectiveness of inputs. Overall, we found few school "inputs" that were associated with differential achievement over time. Frequent use of textbooks increased achievement by a full point on the posttest, while use of workbooks decreased achievement by a third of a point; an enriched curriculum increased posttest scores by over 2.5 points. Each additional percentage point of teachers qualified to teach mathematics raised posttest scores by over 1 point.

However, these causal statements do not hold if they are to be interpreted as the result of an external intervention. Obtaining (additional) textbooks for the schools is not a simple procedure unrelated to educational processes and management decisions; it is itself an outcome variable related to some (unknown) aspects of the educational process. Similarly, discarding workbooks might not lead to improved outcomes, unless all the circumstances that lead to reduced use of workbooks are also present or are induced externally. External intervention will be free of risk only if we have, and apply, causal models for how the educational system functions. The models developed in this paper, and elsewhere in the literature on educational

research, are purely descriptive. Use of regression methods and of variance component analysis allows improved description but does not provide inferences about causal relationships.

In addition, interpretations of the estimates of effects are subject to a variety of influences, and there may be alternative regression models, with different variables, that are equally correct in terms of prediction. Thus, the selection of variables included in this model is responsible, to some degree, for the results, and a different selection of variables could yield substantially different results with respect to the contribution of each variable.

Comparison with OLS. The analysis demonstrates that estimates based on OLS regressions do yield different results, in some cases, from those based on VC regressions. For example, in comparing the OLS estimates with the VCS estimates in Figure 6, we see that for tmthsub the coefficients are quite different. Based on OLS, we would conclude that students in "enriched" classes, with the other explanatory variables controlled for, perform about 2 points (13%) higher than those in "normal" or "remedial" classes; the conclusion based on the VC regression is that they perform nearly 2.6 points (17%) higher. Combining these effects with cost information permits an estimation of cost- effectiveness. If enriched classes cost 13% more than remedial or normal classes, we would conclude that they were either equally cost-effective (OLS) or more cost-effective (VC) than are remedial/normal classes, depending on the model. Similarly, if enriched classes cost 17% more than remedial/normal classes, they would be either equally cost-effective (VC) or less cost-effective (OLS), depending on the model.

However, the caution in the previous subsection about causal inference applies equally in this context. Classes, or schools, cannot be declared to have enriched curricula at an external will and by supplying the outward signs of having enriched curriculum; rather, a whole complex of related circumstances has to be arranged, e.g., strengthened education in lower grades, synchronization with other subjects, etc. Since we argued earlier in the paper that estimates based on VC methods are preferable to those based on OLS methods, differences of these types could hold important policy implications for schools deciding on the type of curriculum to choose.

Caveats

We have noted that alternative models can yield similar predictions (in terms of achievement) but might include a different set of variables. That such could be the case is not a problem limited to VC models; it is a perennial problem with these general types of analyses. In our analysis, we included a number of individual pupil and school/classroom variables; in this respect, we moved well beyond earlier models, which included only modest "intake" characteristics of students. Identifying the variables associated with higher outcome scores does not, however, offer a direct answer to the principal question of a development agency about the distribution of its resources to a set, or continuum, of intervention policies in an educational system. Without any prior knowledge of the educational system, any justification for an intervention policy based on the results of regression (or variance component) analysis, or even of structural modelling (LISREL), has no proper foundation. Certain intervention policies may cause a change in the educational system, and hence a change in the regression model itself.

This new regression model may indicate that the selected intervention is far from optimal or may even be detrimental.

A case in point is the pretest score XROT. Its coefficient is positive and of substantial magnitude. A conceivable intervention policy to raise the XROT scores would be, for example, to provide coaching prior to administering the pretest. Clearly such an intervention, if effective, could lead to a change in the regression formula. Alternatively, if coaching took place between the pretest and posttest, the regression formula would again be changed, but differently. Any number of different scenarios is easy to construct, in which the coefficient on XROT would be close to 1 or substantially lower than .62 (the level obtained in our analysis).

Similarly, indiscriminant reduction of the time spent maintaining order in the classroom, probably a less expensive intervention in monetary terms, is likely to be an unreasonable solution. Introduction of the enriched mathematics curriculum for all students is most likely not practical, and even its extension to a few more classrooms may place excessive requirements on staff in the schools that would lower the quality of instruction in other subjects and/or other grades.

In conclusion, positive or negative regression coefficients cannot be regarded uncritically as indicators of cause and effect, or influence. An intervention should be regarded as an experiment, whose outcome can be predicted from an observational study only under the unrealistic assumptions of the regression formula describing accurately the mechanics of a rigid educational process.

This finding does not mean that absolutely no inferences can be made without a carefully designed experiment. It means that the results of the statistical analysis based violated assumptions of randomization should be supplemented with external information about the complex selection processes and other sources of bias. This adjustment does not submit to a rigorous treatment, and therefore we can only speculate how different our results would have been had we carried out a (hypothetical) experiment instead of a survey.

Three important items of information would assist in answering the question about the allocation of resources:

(i) What are the feasibility and cost of various interventions

(ii) How an intervention will affect other explanatory variables and which aspects of the educational process will remain unaltered after the intervention

(iii) How directly manipulable the "interventions" are.

It is critical to distinguish between the variables that are manifest (unchangeable, e.g., pupil background), that are manipulable (e.g., time spent on a task of a particular kind) and that are manipulable only by direct intervention. For example, the time spent maintaining discipline is a manipulable variable, but it can be manipulated either indirectly (e.g., by making the curriculum more interesting or by providing more suitable or more interesting textbooks) or directly (by changing teacher behavior so as to

ignore disruptive student behavior). Considerations as to effective education policy require attention to directly manipulable variables. In the present analysis, these are the qualifications of the mathematics teachers and the use of textbooks.

REFERENCES

Aitkin, M., & N. Longford. (1986). Statistical modelling issues in school effectiveness studies (with discussion). Journal of the Royal Statistical Society, Series B, 149:1-43.

Avalos, B., & Haddad, W. (1981). A review of teacher effectiveness research. Ottawa: International Development Research Centre.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., & Congdon, Jr., R.T. (1986). An Introduction to HLM: Computer Program and User's Guide, Chicago, University of Chicago (processed)

Coleman, J., Hoffer, T., & Kilgore, S. (1982). High School Achievement: Public, Catholic and Private Schools Compared. New York: Basic.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Dempster, A.P., Rubin, D.B. & Tsutakawa, R.K. (1981). Estimation in covariance component models. Journal of the American Statistical Association, 76, 341-353.

Fuller, B. (1987). Raising school quality in developing countries: What investments boost learning? Review of Educational Research, 57, 255-291.

Goldstein, H. (1984). The methodology of school comparisons. Oxford Review of Education, 10, 69-74.

_____. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika, 73, 43-56.

_____. (1987). Multilevel Models in Educational and Social Research, New York: Oxford University Press.

Harbison R. & E. Hanushek. (1988). Educational performance of the poor: Lessons from rural northeast Brazil. Draft manuscript.

Heyneman, S.P. (1986). The search for school effects in developing countries: 1966-1986. EDI Seminar Paper No. 33. Washington, D.C.: World Bank.

Heyneman, S.P. & Jamison, D.T. (1980). Student learning in Uganda: Textbook Availability and other factors, Comparative Education Review, 23, 206-220

Heyneman, S.P. & Loxley, W. (1983) The effect of primary school quality on academic achievement across twenty-nine high and low-income countries, American Journal of Sociology, 88, 1162-1194.

Husen, T., Saha, L., & Noonan, R. (1978). Teacher training and student achievement in less developed countries (Staff Working Paper 310). Washington DC: The World Bank.

Lindley, D.V. & Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society, Series B, 43, 1-41.

Lockheed, M.E., Fuller, B. & Nyirongo, R. (1987). Family effects on student achievement in Thailand and Malawi, World Bank: Population and Human Resources Department (processed).

Lockheed, M.E., Vail, S. & Fuller, B. (1987) How textbooks affect achievement in developing countries: Evidence from Thailand. Educational Evaluation and Policy Analysis, 8, 379-392

Lockheed, M.E. & Hanushek, E. (1988) Improving educational efficiency in developing countries: What do we know? Compare, 18, 21-38.

Lockheed, M.E., Foncier J. & Bianchi, L. (1989). Effective primary level science teaching in the Philippines. "Paper presented at the annual meeting of the American Sociological Association in San Francisco, California, Aug. 9-13, 1989

Lockheed, M.E. & Komenan, A. (1989). Teaching quality and student achievement in Africa: the case of Nigeria and Swaziland. Teaching + teacher education. Great Britain: Pergamon Press.

Longford, N.T. (1986). VARCL—Interactive software for variance component analysis. The Professional Statistician, 5, 28-32.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika, 74, 817-827.

Mason, W. M., Wong, G. Y. & Entwistle, B. (1984). The multilevel linear model: A better way to do contextual analysis. Sociological Methodology. London: Jossey-Bass.

Psacharopoulos, G. & Loxley, W. (1986). Diversified Secondary Education and Development, Baltimore, MD: Johns Hopkins University Press.

Raudenbush, S.W. (1987). Educational applications of hierarchical linear models: A review. Michigan State University (processed).

Raudenbush, S.W. & Bryk, A.S. (1986). A hierarchical model for studying school effects. Sociology of Education, 59, 1-17

Reynolds, D. (1985). Introduction: Ten years on—a decade of research and activity in school effectiveness research reviewed. In D. Reynolds (Ed.) Studying School Effectiveness. London: The Falmer Press.

Rosenholtz, S. (1989). Teachers' workplace: The social organization of schools. White Plains, NY: Longman. Sirotnik, K., & Burstein, L. (1985).

Rutter, M. (1983). School effects on pupil progress: Research findings and policy implication. Child Development, 54, 1-29.

Schiefelbein, E., & Simmons, J. (1981). Determinants of school achievement: A review of research for developing countries (mimeo). Ottawa: International Development Research Centre.

Sirotnik, K.A. & Burstein, L. (1985). Measurement and statistical issues in multilevel research on schooling. Educational Administration Quarterly, 21, 169-185.

Willms, J.D. (1987). Differences between Scottish education authorities in their examination attainment. Oxford Review of Education, 13, 211-231.

PPR Working Paper Series

	Title	Author	Date	Contact for paper
WPS223	Overvalued and Undervalued Exchange Rates in an Equilibrium Optimizing Model	Jose Saul Lizondo	August 1989	R. Luz 61588
WPS224	The Economics of the Government Budget Constraint	Stanley Fischer	May 1989	S. Fischer 33774
WPS225	Targeting Assistance to the Poor: Using Household Survey Data	Paul Glewwe Oussama Kanaan	June 1989	B. Rosa 33751
WPS226	Inflation and the Costs of Stabilization: Country Experiences, Conceptual Issues, and Policy Lessons	Andres Solimano	July 1989	E. Khine 61763
WPS227	Institutional Reforms in Sector Adjustment Operations	Samuel Paul	July 1989	E. Madrona 61712
WPS228	Recent Economic Performance of Developing Countries	Robert Lynn F. Desmond McCarthy	July 1989	M. Divino 33739
WPS229	The Effect of Demographic Changes on Saving for Life-Cycle Motives in Developing Countries	Steven B. Webb Heidi S. Zia	July 1989	E. Khine 61765
WPS230	Unemployment, Migration, and Wages in Turkey, 1962-85	Bent Hansen	July 1989	J. Timmins 39248
WPS231	The World Bank Revised Minimum Standard Model: Concepts and Issues	Doug Addison	May 1989	J. Onwuemene-Kocha 61750
WPS232	Women and Food Security in Kenya	Nadine R. Horenstein	June 1989	M. Villar 33752
WPS233	Public Enterprise Reform in Adjustment Lending	John Nellis	August 1989	R. Malcolm 61708
WPS234	A Consistency Framework Macroeconomic Analysis	William Easterly	June 1989	R. Luz 61760
WPS235	Borrowing, Resource Transfers, and External Shocks to Developing Countries: Historical and Counterfactual	Steven B. Webb Heidi S. Zia	July 1989	E. Khine 61765
WPS236	Education and Earnings in Peru's Informal Nonfarm Family Enterprises	Peter Moock Philip Musgrove Morton Stelcner	July 1989	M. Fisher 34819

PPR Working Paper Series

	Title	Author	Date	Contact for paper
WPS237	The Curricular Content of Primary Education in Developing Countries	Aaron Benavot David Kamens	June 1989	C. Cristobal 33640
WPS238	The Distributional Consequences of a Tax Reform On a VAT for Pakistan	Ehtisham Ahmad Stephen Ludlow	August 1989	A. Bhalla 60359
WPS239	The Choice Between Unilateral and Multilateral Trade Liberalization Strategies	Julio Nogues	July 1989	S. Torrijos 33709
WPS240	The Public Role in Private Post-Secondary Education: A Review of Issues and Options	Ake Blomqvist Emmanuel Jimenez	August 1989	A. Bhalla 61059
WPS241	The Effect of Job Training on Peruvian Women's Employment and Wages	Ana-Maria Arriagada	July 1989	C. Cristobal 33640
WPS242	A Multi-Level Model of School Effectiveness in a Developing Country	Marlaine E. Lockheed Nicholas T. Longford	July 1989	C. Cristobal 33640
WPS243	Averting Financial Crisis - Kuwait	Fawzi H. Al-Sultan	July 1989	R. Simaan 72167
WPS244	Do Caribbean Exporters Pay Higher Freight Costs?	Alexander J. Yeats	July 1989	J. Epps 33710
WPS245	Developing a Partnership of Indigenous Peoples, Conservationists, and Land Use Planners in Latin America	Peter Poole	August 1989	S. Davis 38622
WPS246	Causes of Adult Deaths in Developing Countries: A Review of Data and Methods	Richard Hayes Thierry Mertens Geraldine Lockett Laura Rodrigues	July 1989	S. Ainsworth 31091
WPS247	Macroeconomic Policies for Structural Adjustment	Carlos Alfredo Rodriguez	August 1989	R. Luz 61588
WPS248	Private Investment, Government Policy, and Foreign Capital in Zimbabwe	Mansoor Dailami Michael Walton	August 1989	M. Raggambi 61696
WPS249	The Determinants of Hospital Costs: An Analysis of Ethiopia	Ricardo Bitran-Dicowsky David W. Dunlop	August 1989	V. Israel 48121
WPS250	The Baker Plan: Progress, Shortcomings, and Future	William R. Cline	August 1989	S. King-Watson 33730